

MODÉLISATION,
ENRICHISSEMENT SÉMANTIQUE
ET DIFFUSION D'UN CORPUS
TEXTUEL SEMI-STRUCTURÉ : LE
CAS DES CATALOGUES DE VENTE
DE MANUSCRITS.

Paul, Hector Kervegan

25 septembre 2022

En quoi la nature semi-structurée du corpus permet d'en automatiser le traitement ? Comment produire des informations normalisées et exploitables à partir d'un corpus textuel semi-structuré ? Comment rendre la recherche en humanités numériques réutilisable et encourager le partage de données entre projets de recherche ?

Introduction

Plan

- 1 La structure du texte comme méthode d'approche
 - Structure du document et modélisation
 - La spécificité d'un corpus « semi-structuré »
- 2 Sous quels angles aborder cette problématique ?
 - Modéliser un corpus semi-structuré
 - Analyser le texte à partir de sa structure : la résolution d'entités nommées
 - Remodeler, re-modéliser le texte pour une API
- 3 Le traitement automatique du texte comme chaîne éditoriale continue
 - L'enrichissement progressif du texte
 - La résolution d'entités nommées : mettre le texte en réseau
 - Recomposer le texte et créer des documents via une API

- 1 La structure du texte comme méthode d'approche
 - Structure du document et modélisation
 - La spécificité d'un corpus « semi-structuré »
- 2 Sous quels angles aborder cette problématique ?
 - Modéliser un corpus semi-structuré
 - Analyser le texte à partir de sa structure : la résolution d'entités nommées
 - Remodeler, re-modéliser le texte pour une API
- 3 Le traitement automatique du texte comme chaîne éditoriale continue
 - L'enrichissement progressif du texte
 - La résolution d'entités nommées : mettre le texte en réseau
 - Recomposer le texte et créer des documents via une API

Étudier la structure du texte permet de faire le lien entre un document physique et son édition numérique.

19. Commune de 1871.

46 lettres ou pièces originales de personnages ayant coopéré à la Commune de 1871, environ 100 p. in-4 ou in-8.

Important dossier historique. Parmi les noms qui y sont contenus on remarque les suivants: *Grousset* (Paschal), *Longuet*, *Ranvier*, *Lullier*, *Miot*, *Michel* (Louise), *Bergeret*, *Cluseret*, *Malon*, *Lefrançais*, *Delescluze*, *Urbain*, *Tridon*, *Vermorel*, etc.

FIGURE – Un manuscrit vendu aux enchères

La structure du texte comme méthode d'approche

Structure du document et modélisation

```
1 <item n="19" xml:id="CAT_000193_e19">
2   <num type="lot">19</num>
3   <name type="author" ref="wd:Q63314476">Commune de 1871</name>
4   <desc xml:id="CAT_000193_e19_d1">
5     <date when="1871">46 lettres ou pièces originales de personnages ayant coopéré à la
6       ↳ Commune de 1871</date>, environ <measure type="length" unit="p" n="100">100
7       ↳ p.</measure><measure type="format" unit="f" ana="#document_format_4">in-4
8       ↳ o</measure>u in-8
9   </desc>
10  <note>Important dossier historique. Parmi les noms qui y sont contenus on remarque les
11    ↳ suivants: Grousset (Paschal), Longuet, Ranvier, Lullier, Miot, Michel (Louise),
12    ↳ Bergeret, Cluseret, Malon, Lefrançais, Delescluze, Urbain, Tridon, Vermorel,
13    ↳ etc.</note>
14 </item>
```

DONNÉES BRUTES 1 – L'encodage du même manuscrit en TEI

La structure du texte comme méthode d'approche

La spécificité d'un corpus « semi-structuré »

« Semi-structuré », qu'est-ce que ça veut dire ?

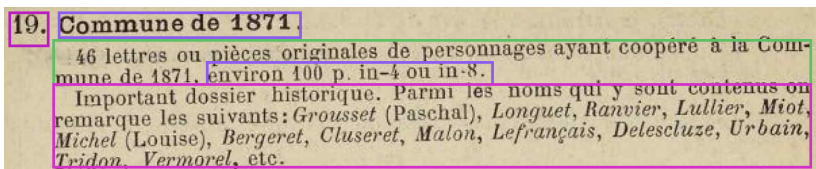


FIGURE – Les différents éléments dans la description du manuscrit

La structure du texte comme méthode d'approche

La spécificité d'un corpus « semi-structuré »

L'édition critique, un autre type de structure formelle pour le texte

Ces cinq seigneurs pour venir devers le conte de Fois et requier en nom de mariage pour le duc de Berry ceste jenne dame , se partirent de leurs lieux et se devoient trouver en Avignon deléz le pape comme ils firent. Ce
35 pape Clement, quy cousin estoit du pere a la damoiselle, les retint la bien quinze jours. Et environ la Chandelleur ilz se departirent tous d' Avignon et prindrent le chemin de Nismes et de Montpellier. Si chevauchierent a petites journees et a grans frais. Si passerent Besiers et vindrent a Carcassonne.

17 trop] ϕ ATW BRN || 17 luy] lui ATW BRN || 18 ,] ϕ ATW BRN || 20 d'Arneignach.
Et] d'Armignac et ATW; de Armignac BRN || 21-22 , voire si longuement] tant ATW
BRN || 22 parfaite et] parfaite BRN || 22 fourmee] formee ATW; ϕ BRN || 22
respondy] dit ATW; respondit BRN || 24 "Or ça, beaulx oncles] "Bel oncle ATW BRN
|| 24 si tres] tant ATW; si BRN || 25] , ATW || 25 moult] ϕ ATW BRN || 26 gaires]
guerres ATW || 27 Burel] Brueil ATW BRN || 27 chevalier] ϕ ATW BRN || 27 et] ϕ
BRN || 27 chambellan] chambrelan ATW; chambrelain BRN || 27 ,] ϕ ATW || 28 .
Et avecques] et avecques ATW BRN || 30 et] . Et ATW BRN || 30 et] ϕ ATW || 31
ung] , un BRN || 31 , sage] saige ATW BRN || 32] pour ATW || 33 de mariage]
|| 33] a mariage ATW || 33 ,] ϕ BRN || 36 tous] ϕ ATW BRN || 38 frais. Si] fraiz
et ATW BRN || 38-39 . Et la] ou ATW; et BRN

FIGURE – Extrait d'une édition critique de trois témoins du SHF 3A-306 « La négociation du mariage du Duc de Berry » des *Chroniques* de Froissart.

- 1 La structure du texte comme méthode d'approche
 - Structure du document et modélisation
 - La spécificité d'un corpus « semi-structuré »
- 2 Sous quels angles aborder cette problématique ?
 - Modéliser un corpus semi-structuré
 - Analyser le texte à partir de sa structure : la résolution d'entités nommées
 - Remodeler, re-modéliser le texte pour une API
- 3 Le traitement automatique du texte comme chaîne éditoriale continue
 - L'enrichissement progressif du texte
 - La résolution d'entités nommées : mettre le texte en réseau
 - Recomposer le texte et créer des documents via une API

Sous quels angles aborder cette problématique ?

Modéliser un corpus semi-structuré

Modéliser un document revient à identifier sa structure et à l'explicitier à l'aide d'une structure formelle explicite.

```
1 <item n="19" xml:id="CAT_000193_e19">
2   <num type="lot">19</num>
3   <name type="author" ref="wd:Q63314476">Commune de 1871</name>
4   <desc xml:id="CAT_000193_e19_d1">
5     <date when="1871">46 lettres ou pièces originales de personnages ayant coopéré à la
        ↳ Commune de 1871</date>, environ <measure type="length" unit="p" n="100">100
        ↳ p.</measure><measure type="format" unit="f" ana="#document_format_4">in-4
        ↳ o</measure>u in-8
6   </desc>
7   <note>Important dossier historique. Parmi les noms qui y sont contenus on remarque les
        ↳ suivants: Grousset (Paschal), Longuet, Ranvier, Lullier, Miot, Michel (Louise),
        ↳ Bergeret, Cluseret, Malon, Lefrançais, Delescluze, Urbain, Tridon, Vermorel,
        ↳ etc.</note>
8 </item>
```

Sous quels angles aborder cette problématique ?

Modéliser un corpus semi-structuré

Les limites de la modélisation

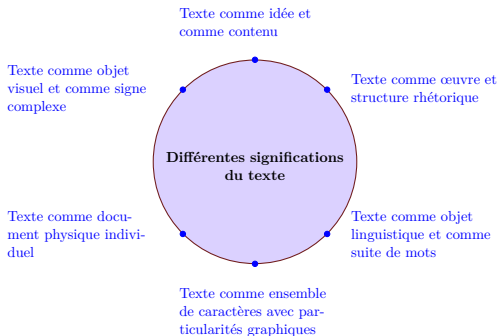


FIGURE – Les différentes significations d'un texte selon P. Sahle (2016)

Sous quels angles aborder cette problématique ?
Analyser le texte à partir de sa structure

La résolution d'entités nommées à partir de détection de motifs



nom de famille noble (prénom de la nom de famille usuel titre de noblesse de)

FIGURE – Les différents motifs dans un nom propre

Sous quels angles aborder cette problématique ?

Remodeler, re-modéliser le texte pour une API

Entre texte et données, les catalogues de vente sont des sources utiles pour constituer une API :

- Les descriptions de manuscrits sont quelque part entre texte et donnée
- Ils sont composés de plusieurs entités significantes en elles-mêmes
- Constituer une API permet de créer de nouveaux points d'entrée dans les catalogues

Plan

- 1 La structure du texte comme méthode d'approche
 - Structure du document et modélisation
 - La spécificité d'un corpus « semi-structuré »
- 2 Sous quels angles aborder cette problématique ?
 - Modéliser un corpus semi-structuré
 - Analyser le texte à partir de sa structure : la résolution d'entités nommées
 - Remodeler, re-modéliser le texte pour une API
- 3 Le traitement automatique du texte comme chaîne éditoriale continue
 - L'enrichissement progressif du texte
 - La résolution d'entités nommées : mettre le texte en réseau
 - Recomposer le texte et créer des documents via une API

Une chaîne éditoriale continue

L'enrichissement progressif du texte

Contrairement à une édition papier, une édition numérique est vouée à être continuellement enrichie.

```
1 <item n="19" xml:id="CAT_000193_e19">
2   <num type="lot">19</num>
3   <name type="author">Commune de 1871</name>
4   <desc xml:id="CAT_000193_e19_d1">46 lettres ou pièces originales de personnages ayant
   ↳ coopéré à la Commune de 1871, environ 100 p. in-4 ou in-8</desc>
5   <note>Important dossier historique. Parmi les noms qui y sont contenus on remarque les
   ↳ suivants: Grousset (Paschal), Longuet, Ranvier, Lullier, Miot, Michel (Louise),
   ↳ Bergeret, Cluseret, Malon, Lefrançais, Delescluze, Urbain, Tridon, Vermorel,
   ↳ etc.</note>
6 </item>
```

DONNÉES BRUTES 3 – Une description de manuscrit au début
de la chaîne de traitement

Une chaîne éditoriale continue

L'enrichissement progressif du texte

```
1 <item n="19" xml:id="CAT_000193_e19">
2   <num type="lot">19</num>
3   <name type="author" ref="wd:Q63314476">Commune de 1871</name>
4   <desc xml:id="CAT_000193_e19_d1">
5     <date when="1871">46 lettres ou pièces originales de personnages ayant coopéré à la
6     ↳ Commune de 1871</date>, environ <measure type="length" unit="p" n="100">100
7     ↳ p.</measure><measure type="format" unit="f" ana="#document_format_4">in-4
8     ↳ o</measure>u in-8
9   </desc>
10  <note>Important dossier historique. Parmi les noms qui y sont contenus on remarque les
11  ↳ suivants: Grousset (Paschal), Longuet, Ranvier, Lullier, Miot, Michel (Louise),
12  ↳ Bergeret, Cluseret, Malon, Lefrançais, Delescluze, Urbain, Tridon, Vermorel,
13  ↳ etc.</note>
14 </item>
```

DONNÉES BRUTES 4 – La même description à la fin de la chaîne de traitement

Une chaîne éditoriale continue

La résolution d'entités nommées : mettre le texte en réseau

La résolution d'entités nommées permet de connecter le texte à d'autres données dans une logique d'hypertexte.

Une chaîne éditoriale continue

Recomposer le texte et créer des documents via une API

```
1 <publicationStmt>
2   <ab>
3     <date type="file-creation-date"
4       ↪  when-iso="2022-08-24T16:41:40.267860">2022-08-24T16:41:40.267860</date>
5       <ref type="http-status-code"
6         ↪  target="https://developer.mozilla.org/en-US/docs/Web/HTTP/Status/200">200</ref>
7       <note>Original data made available under Creative Commons Attribution 2.0 Generic (CC BY
8         ↪  2.0)</note>
9       <table>
10        <head>Query parameters</head>
11        <row><!-- en-têtes --></row>
12        <row>
13          <cell role="value" corresp="level">item</cell>
14          <cell role="value" corresp="id">CAT_000204_e108_d1</cell>
15          <cell role="value" corresp="format">tei</cell>
16        </row>
17      </table>
18    </ab>
19  </publicationStmt>
```

Comment la modélisation et l'éditorialisation continue du texte impactent celui-ci ?

- L'encodage comme « métatexte »
- L'API et les limites d'une conception traditionnelle du texte